# A Survey of Domain Specific Web Search Techniques

Vidya Vadke
M.E (Computer) Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Pune, India
Savitribai Phule Pune University, Pune, Maharashtra, India -411007

Aruna Gupta
Assistant Professor Department of IT, Jayawantrao Sawant College of Engineering, Pune, India
Savitribai Phule Pune University, Pune, Maharashtra, India -411007

## ABSTRACT
In the world of internet where every day information is increasing exponentially, retrieving correct information from the World Wide Web has always remained a challenge. The growth in volume of data has made it more difficult to find relevant and useful information on the internet. It is always a challenge to get relevant information if it is searched over the vast internet. There are techniques to crawl the domain specific data so as to keep the throw away data as to minimum. Taking the advantage of knowing domain, priorities could be assigned to the indexes for faster search. Assistance could be given to the user to form the most effective query in the given domain. Some search techniques make use of URL structure, some look for specific words in the web pages. Number of attempts have been made to implement domain specific search namely by combining semantic web technologies with information retrieval. Making use of weighted ontology for crawling, indexing and searching has been taken into consideration by some of researchers separately. In the need to improve user experience in domain specific information retrieval, this is an attempt to review the techniques available for domain specific web crawling, indexing and searching. The aim is to focus on the areas which can be improved to give enriching search experience to the users of domain specific search.

## General Terms
Adaptive ontology, Domain specific search.

## Keywords
Domain Specific web search, Search engine, Ontology, indexing, web crawler.

## 1. INTRODUCTION
To be able to extract relevant information from Internet, search engines are used. Popular and widely used search engines are generic and try to cover all the domains to be able to serve all. In the attempt, user is presented with the information in all domains which have relevance to the user query.

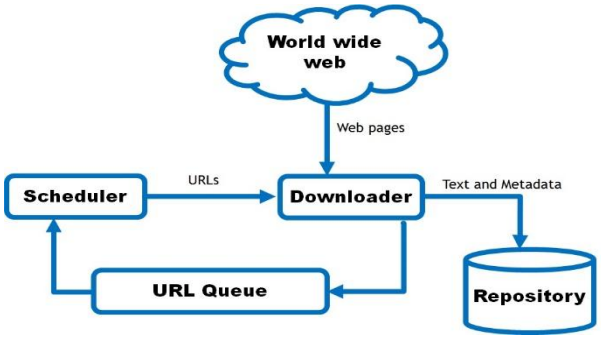


**Fig 1: Generic web crawler architecture.**

Typical web crawler architecture is shown in the Figure 1. Now a days most of the popular search engines make use of cloud based web crawlers to deal with humongous data they need to crawl and gather in their repository.

If the user is aware of the domain then he prefers to go for the sites which provide his domain specific information rather than searching for relevant results in the generic search engine. In that way, it is guaranteed that only information related to relevant domain is retrieved. Domain specific search engines guarantee that users will get information only from the relevant domain.

Domain specific crawling is popularly known as focused crawling. Focused crawlers fetch Web documents selectively that are relevant to a specified topics.

Though the volume of data that need to be crawled, indexed and searched is lesser in domain specific search as compared to generic search engines, domain specific search engines need to be more intelligent while crawling so as to get data relevant to specific domain; while indexing so as to prioritize the index terms and while searching so as to assist the user to form the effective query.

Several search engines using focused web crawler have been generated in recent years. The topic-specific web crawler fetches web pages from the Internet by gathering only specific pages related to a domain and needs smaller storage, they could offer higher precision and recall ratio than generic search engines.

Seed URLs used in domain specific web crawler play an important role. As for gathering domain specific information, the relevant sites are to be chosen in the first place to reduce the crawling wastage.

Sometimes domain terminologies are grouped into domain dimensions and then these domain dimensions are used to index the documents crawled and to develop the search interface which could be made interactive based on the dimensions. Sometimes the domain dimensions are represented in form of Ontology. Ontologies could be used to decide whether or not to add the web pages to the repository.

In generic crawlers, indexes are creating by parsing the web pages indexing all the words and phrases. Stop words are filtered to reduce the noise. In domain specific search engines, indexes are generally predefined terms based on the domain knowledge. Index terms can also be prioritized based on the weightage of a particular term with respect to the domain.

The generic search engine can provide assistance to user while user is forming the query based on the popularity and similarity of the query string user is typing; whereas domain specific search engines can provide interactive means for user to be able to extract the most relevant information in the specific domain.

## 2. LITERATURE SURVEY AND RELATED WORK

Various efforts have been taken by researchers to improve the efficiency of focused crawling. Web pages can be characterized by the hypertext. The hypertext can be used to check the relevance of the document to the domain. The semantics of the link specifies the semantics of the web page.

LSCrawler proposed in [1] retrieves web pages by inspecting the relevancy of the page based on the words in the link and the surrounding text of the link. The relevancy of the page is calculated by measuring the semantic similarity between the words in the link and the taxonomy hierarchy of the given domain. Though it enhances the process of determining the relevancy of the documents before downloading; it would have performed better if a rule based inference system was developed; as the keywords in hypertext might not always describe the concept correctly.

UBFC (URL rule Based Focused Crawler) [2] algorithm based on a double-crawler (an experimental crawler and a focused crawler. The heart of the UBFC is an URL regular expression learner, which automatically learns and generalizes the regular expressions of URLs of the web pages. Though the harvest rate and recall rate are better using UBFC; it needs to be fed the list of domain-related websites manually.

Ontology-based Web Crawler [3] proposes a new metric called Association metric that is based on the semantic content of the URL and all its parent pages along with the importance metric. It proposes a new method for prioritizing the URL queue for crawling by taking account both the semantic and link structure of the web. It only works around assigning priority to the URLs in frontier queue.

Another way suggested for domain specific crawling using ontology is as per suggested by [4]. Ontology is used to determine whether to add a traversed page in the repository or not. Its approach is surely an improvement over the approaches suggested in [1], [2], [3] as it actually traverses the web page to find out the relevance and does not depend only on URL structure, link semantics etc. But the major disadvantage of the approach proposed in [4] is that it needs to actually download and traverse all the pages to decide whether the page is relevant or not; which makes the network bandwidth usage similar to generic crawler.

The concept of using detailed ontology for web crawling is further extended in [4] by applying priority in Ontology terms to web-page indexing. Web pages are index based on multiple levels of dominating and sub-dominating ontology terms appearing in web pages. Results of [4] are by measured by providing drop-down list of ontology terms which are treated as search query. This approach deprives user from forming his own query. Also the search functionality is limited by selected drop-down terms and user not being able to form his own query.

Adaptive Focused Crawling Based on Link Analysis proposed in [8] calculates the unvisited URL score based on its Anchor text relevancy, its description in Google search engine and calculate the similarity score of description with topic keywords, cohesive text similarity with topic keywords and Relevancy score of its parent pages. Disadvantage of adaptive focused crawler in [5] is that it depends on the text description provided in Google search for calculating text relevancy.

A framework is developed and tested in [11] for adaptive ontology based focused crawling and searching. Use of Ontology is done to represent domain and user profile as well. Though [11] has put together efforts for crawling and searching based on adaptive ontology, there are certain shortcomings like users are classified to very broad categories, users are asked to select the adaptive ontology and his profile characteristics which is only one time, users are boxed to the profile chosen for the first time and the domain ontology too which itself does not cater to the adaptive nature. This framework too heavily depends on Google search for crawling and searching too. Also the document base used for the experiments is very limited which limits the possibility of arriving at real time results.

## 3. CONCLUSION

In the need of having better domain specific search and user experience, various techniques related to domain specific web crawling, indexing and searching are reviewed. Number of researchers have focused on specific areas of domain specific web search. Some efforts are made towards considering domain specific web search as a homogeneous system and developing frameworks as a whole. Though adaptive ontology based approach is a promising approach; there is scope for more research in improving domain specific document corpus, user's search experience, query formation and processing leveraging the domain knowledge and ontology. Existing search engines can be used in more constructive way to help enrich the document base.

## 4. REFERENCES

[1] M.Yuvarani, N.Ch.S.N.Iyengar, A.Kannan, "LSCrawler: A Framework for an Enhanced Focused Web Crawler based on Link Semantics", Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)

[2] Xiaolin Zheng, Tao Zhou, Zukun Yu,Deren Chen, "URL Rule Based Focused Crawlers", International Conference on e-Business Engineering IEEE DOI 10.1109/ICEBE.2008.61

[3] S.Ganesh, M.Jayaraj, V.Kalyan Srinivasa Murthy, G.Aghila, "Ontology-based Web Crawler", Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC04) 0-7695-2108-8/04 2004 IEEE

[4] Mukhopadhyay D., Biswas A., Sinha, "Web-Page Indexing Based on the Prioritized Ontology Terms", Advanced Computing, Networking and Informatics - Volume 1, Smart Innovation, Systems and Technologies 27, Springer International Publishing Switzerland 2014

[5] Mukhopadhyay D., Biswas A., Sinha, "A New Approach to Design Domain Specific Ontology Based Web Crawler", 10th International Conference on Information Technology, 2007 IEEE pp. 289291

[6] Wei Li, Debasis Ganguly, Gareth J. F. Jones, "Enhanced Information Retrieval Using Domain-Specific Recommender Models", Chapter in Advances in Information Retrieval Theory Volume 6931 of the series Lecture Notes in Computer Science pp 201-212

[7] "OWL Ontology reference", http://www.w3.org/TR/owl-ref/

[8] Debashis Hati, Biswajit Sahoo, Amritesh Kumar, "Adaptive Focused Crawling Based on Link Analysis",

2nd International Conference on Education Technology and Computer (ICETC) 2010 IEEE

[9] Yuekui Yang, Yajun Du, Yufeng Hai, Zhaoqiong Gao, "A Topic-Specific Web Crawler with Web Page Hierarchy Based on HTML Dom-Tree", Asia-Pacific Conference on Information Processing 2009 IEEE DOI 10.1109/APCIP.2009.110

[10] Anirban Kundu, Ruma Dutta, Debajyoti Mukhopadhyay, "An Alternate Way to Rank Hyper-linked Web-Pages", 9th International Conference on Information Technology (ICIT'06) IEEE 2006

[11] Nicolas Guelfi, Cedric Pruski, Chantal Reynaud, "Experimental Assessment of the TARGET Adaptive Ontology-based Web Search Framework", New Technologies of Distributed Systems (NOTERE), 2010 10th Annual International Conference on June 2010 pp 297 - 302